

Regression Models - Assignment

Benil Mathew

22 October 2016

Executive Summary

This report is a R markdown of the peer-graded assignment for Regression Models Coursera course.

The analysis in this assignment attempts to answer the following questions based on the `mtcars` dataset:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions

Based on the analysis MPG changes by $14.0794278 + -4.1413764 * \text{weight}$ (in 1000 lbs) for manual transmission in comparison with automatic transmission, when 1/4 mile time and weight are held constant. Below a weight of 3399.698 lbs manual transmission is better for MPG, but for weight above this value, automatic transmission is better.

The analysis

The `mtcars` dataset include data that was extracted from the 1974 Motor Trend US magazine, comprising of fuel consumption and 10 other aspects of automobile design and performance for 32 automobiles (1973–74 models). The features in the dataset are - `mpg` - Miles/(US) gallon, `cyl` - Number of cylinders, `disp` - Displacement (cu.in.), `hp` - Gross horsepower, `drat` - Rear axle ratio, `wt` - Weight (1000 lbs), `qsec` - 1/4 mile time, `vs` - V/S, `am` - Transmission (0 = automatic, 1 = manual), `gear` - Number of forward gears, `carb` - Number of carburetors

Loading the data

```
library(ggplot2)
data(mtcars)
```

Convert features to factors

```
mtcars$cyl <- as.factor(mtcars$cyl); mtcars$vs <- as.factor(mtcars$vs);
mtcars$am <- as.factor(mtcars$am); mtcars$gear <- as.factor(mtcars$gear);
mtcars$carb <- as.factor(mtcars$carb)
```

Following are the steps followed in identifying the best possible model to identify how `am` affects `mpg`

Plot a pairs plot of key variables that are expected affect `mpg`

1. Determine parameters to include based on what was identified in the pairs plot
2. Run model
3. Assess the p-value for the coefficients, Adjusted R squared and Residual standard error for the degrees of freedom
4. If untried alternate models exist, go to step 1

Based on the assessment, the model chosen include `am`, `wt`, `qsec` and an interaction between `wt` and `am` - `lm(mpg ~ am + wt + qsec + wt*am, data = mtcars)`. Output of the key factors assessed are shown here:

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	9.723053	5.8990407	1.648243	0.1108925394
## am1	14.079428	3.4352512	4.098515	0.0003408693
## wt	-2.936531	0.6660253	-4.409038	0.0001488947
## qsec	1.016974	0.2520152	4.035366	0.0004030165
## am1:wt	-4.141376	1.1968119	-3.460340	0.0018085763

Low p-values for the coefficients (except the Intercept), showing a high level of significance at 0.95 significance level.

Based on the adjusted R squared the model explains 88% of the variance

Residual standard error with this model is 2.0841223 at 27 degrees of freedom

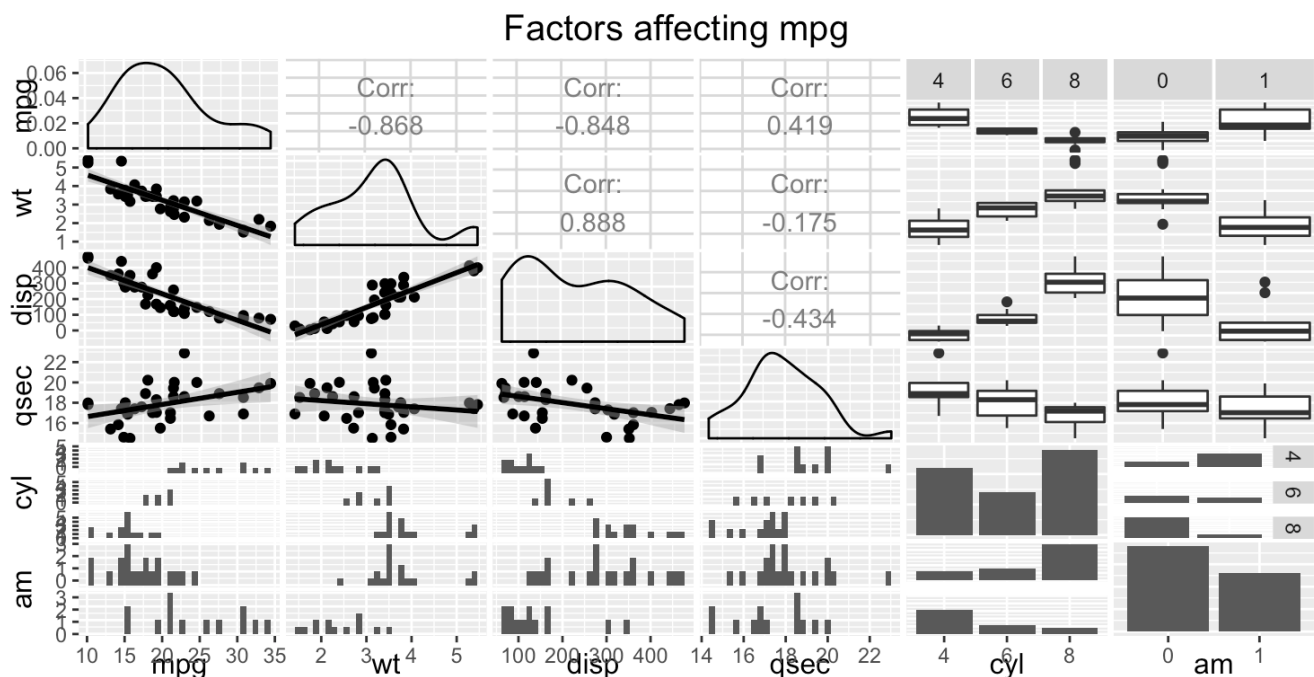
Conclusion

- Is an automatic or manual transmission better for MPG?
 - This will depend on the `wt` of the car. Below `wt` of 3.399698 manual transmission is better for MPG, but for `wt` above this value, automatic transmission is better.
- Quantify the MPG difference between automatic and manual transmissions
 - MPG changes by $14.0794278 + -4.1413764 * wt$ for manual transmission in comparison with automatic transmission, when `qsec` and `wt` are held constant.

Various models that assessed in the process are shown below. They are shown in Appendix along with the outcomes

Appendix

Pair plot of key variables that are expected to impact mpg



Some initial observations, from the plot, that may be relevant in decidnig inclusion or exclusion of features in the model:

- * Both `am` and `cyl` seem to have an effect on `mpg`
- * `mpg` is negatively correlated with `wt` and `disp`, at -0.87 and -0.85 respectively. Inclusion of both in the model may impact the fit, but one of them
- * `mpg` is positively correlated with `qsec`, but is relatively weak at 0.42, a potential candidate for inclusion in the model

- * wt seems to be affected by am, hence may be a factor to be considered in the model
- * Highest correlation among continuous variables is between wt and displacement at 0.89. This may mean that inclusion of both variables in the model is not a good idea

1. Model with only am included

```
fit_am <- lm(mpg ~ am, data = mtcars)
summ <- summary(fit_am)
#summ$coefficients
```

Low p-values for the coefficients show a high level of significance at 0.95 significance level.

Based on the adjusted R squared the model explains 33.8% of the variance

Residual standard error with this model is 4.9020288 at 30 degrees of freedom

2. Attempt a model with all variables

```
fit_all <- lm(mpg ~ ., data = mtcars)
summ <- summary(fit_all)
#summ$coefficients
```

p-values for all coefficients are higher than 0.05, pointing to low significance at 0.95 significance level

Based on the adjusted R squared the model explains 77.9% of the variance

Residual standard error with this model is 2.8331687 at 15 degrees of freedom

3. Adding wt to the model

```
fit_wt <- lm(mpg ~ am + wt, data = mtcars)
summ <- summary(fit_wt)
#summ$coefficients
```

Higher than 0.05 for p-value for the coefficient of covariate am show lower levels of significance at 0.95 significance level.

Based on the adjusted R squared the model explains 73.6% of the variance

Residual standard error with this model is 3.0979398 at 29 degrees of freedom

4. Adding disp to the model

```
fit_wt_disp <- lm(mpg ~ am + wt + disp, data = mtcars)
summ <- summary(fit_wt_disp)
#summ$coefficients
```

No significant improvement, overall, for the p-values of coefficients

Based on the adjusted R squared the model explains 75.8% of the variance

Residual standard error with this model is 2.9674199 at 28 degrees of freedom

5. Replacing disp with qsec to the model

```
fit_wt_qsec <- lm(mpg ~ am + wt + qsec, data = mtcars)
summ <- summary(fit_wt_qsec)
#summ$coefficients
```

Low p-values for the coefficients (except the Intercept) show a high level of significance at 0.95 significance level.

Based on the adjusted R squared the model explains 83.4% of the variance

Residual standard error with this model is 2.4588465 at 28 degrees of freedom

6. Adding cyl as an interaction with wt

```
fit_wt_qsec_int <- lm(mpg ~ am + wt + qsec + wt*cyl, data = mtcars)
summ <- summary(fit_wt_qsec_int)
#summ$coefficients
```

Mixture of p-values for coefficients above and below 0.05

Based on the adjusted R squared the model explains 85.3% of the variance

Residual standard error with this model is 2.3136867 at 24 degrees of freedom

7. Replacing cyl with disp as an interaction with wt

```
fit_wt_qsec_int2 <- lm(mpg ~ am + wt + qsec + wt*disp, data = mtcars)
summ <- summary(fit_wt_qsec_int2)
#summ$coefficients
```

Mixture of p-values for coefficients above and below the threshold of 0.05

Based on the adjusted R squared the model explains 86% of the variance

Residual standard error with this model is 2.2542506 at 26 degrees of freedom

8. Replacing disp with am as an interaction with wt

```
fit_wt_qsec_int3 <- lm(mpg ~ am + wt + qsec + wt*am, data = mtcars)
summ <- summary(fit_wt_qsec_int3)
#summ$coefficients
```

Low p-values for the coefficients (except the Intercept), showing a high level of significance at 0.95 significance level.

Based on the adjusted R squared the model explains 88% of the variance

Residual standard error with this model is 2.0841223 at 27 degrees of freedom

9. Adding another interaction of qsec and cyl

```
fit_wt_qsec_int4 <- lm(mpg ~ am + wt + qsec + wt*am + qsec*cyl, data = mtcars)
summ <- summary(fit_wt_qsec_int4)
#summ$coefficients
```

Mixture of p-values for coefficients above and below the threshold of 0.05

Based on the adjusted R squared the model explains 86.3% of the variance

Residual standard error with this model is 2.2277839 at 23 degrees of freedom

10. Replacing wt in model 3 with disp

```
fit_disp <- lm(mpg ~ am + disp, data = mtcars)
summ <- summary(fit_disp)
#summ$coefficients
```

Mixture of p-values for coefficients above and below the threshold of 0.05

Based on the adjusted R squared the model explains 71.5% of the variance

Residual standard error with this model is 3.2178456 at 29 degrees of freedom

11. Replacing wt in model 8 with disp

```
fit_disp_qsec_int5 <- lm(mpg ~ am + disp + qsec + wt*am, data = mtcars)
summ <- summary(fit_disp_qsec_int5)
#summ$coefficients
```

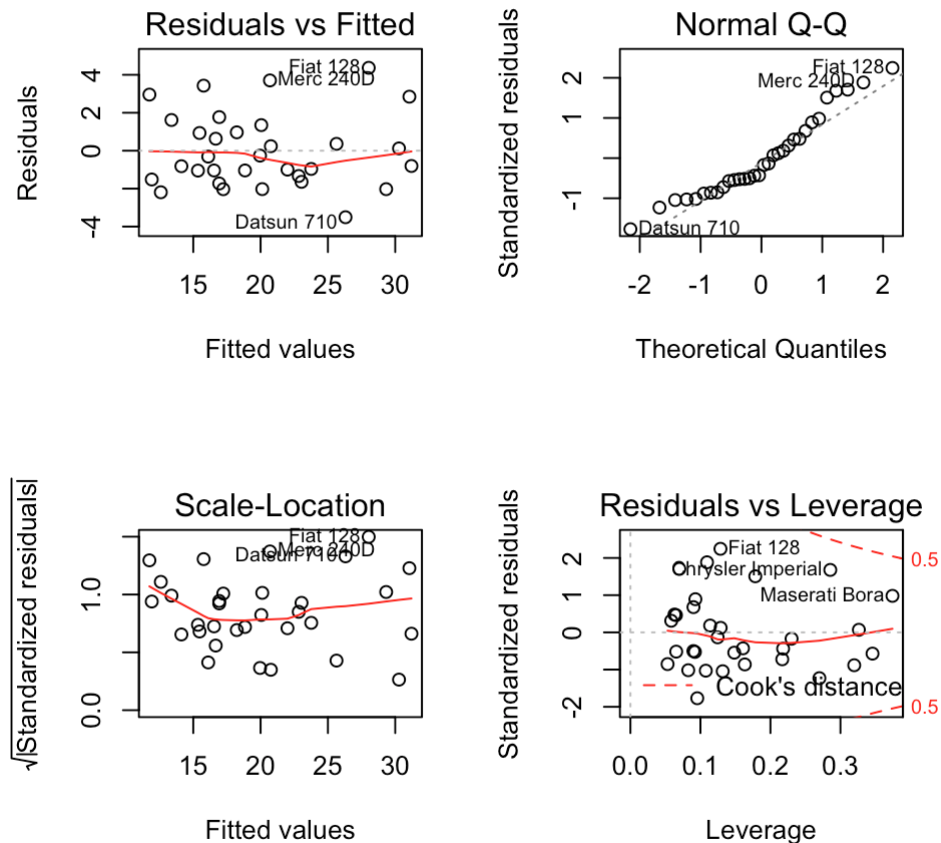
Compared to model 8, p-values appear to show that w_t is more suited in the model

Based on the adjusted R squared the model explains 87.6% of the variance

Residual standard error with this model is 2.1230447 at 26 degrees of freedom

Based on the analysis, model 8 is chosen as the most appropriate. Plotting residuals to confirm that there are no other problems.

Plot of the fit



The plot shows:

- no discernable patterns with residuals plotted against fitted values
- points close to the line in QQ plot, hence close to normally distributed
- no points exertng high leverage as outliers